Quantifying and Neutralising Sexually Explicit Language

George R. S. Weir[1]
Computer & Information Sciences
University of Strathclyde
Glasgow, UK

george.weir@strath.ac.uk

Abstract

Detecting the degree and character of sexually explicit textual content is a feasible and potentially useful enterprise. Such a facility may assist in preventing child exploitation, for example, by automating the detection of the highly sexualised content of on-line grooming, and may be beneficial in SMS for detecting offensive sexting. In addition, an ability to determine the nature and 'strength' of sexually explicit content would prove helpful in contexts where students, trainees or other professionals need to be exposed to sexually explicit language in documents. In such settings, we might deploy such detection and quantification toward managing the content, for instance, by means of progressive 'neutralisation'. In the following, we elaborate upon the relevance of such quantification and describe a variety of steps toward neutralisation.

1. Introduction

The COPINE (Combating Paedophile Information Networks in Europe) classification model [1] rates the severity of victimisation in child pornographic imagery on a ten point scale [2]. In the UK and elsewhere in Europe, this scheme is used as an aid to investigation and criminal proceedings and has been adapted for use by the UK's Sentencing Advisory Panel as a five point scale (Figure 1) for use in court cases [3].

Such classifications afford a standard approach for gauging the seriousness of offences, based upon the 'strength' of the imagery possessed or exchanged by alleged offenders. In court, the scheme can assist in limiting the need to expose individuals - those attending or participating in the proceedings - to graphic, potentially distressing or damaging materials. Instead, an outline

[1] I am grateful for practical insights to several former students (Ana-Marie Duta, Christopher Forbes and Katherine Darroch) whom I have supervised on this topic in the project components of their respective degrees.

description of the classification scheme, combined with the degree and quantity of the materials in a suspect's possession, can convey the significant dimensions in a criminal case, that is, the degree or 'strength' of individual images or video content and the quantity of materials that fall under each classification.

1.	Images depicting nudity or erotic posing with no sexual activity
2.	Sexual activity between children, or solo masturbation by a child
3.	Non-penetrative sexual activity between adult(s) and child(ren)
4.	Penetrative sexual activity between child(ren) and adult(s)
5.	Sadism or bestiality

Fig. 1: SAP scale

For some professionals in law enforcement, legal representation or digital forensics, exposure to hard-core pornography in the form of explicit images and video is often a necessary aspect of the job. Such individuals may be called upon to scrutinise or pass judgement on the materials with which they are confronted. Since they may be required to apply such classification schemes, they will not directly benefit from schemes of this type.

Another major area of concern for law enforcement in which content has a crucial role is the use of social networking and similar chat-based sites for the sexual grooming of minors. Through such networking sites, those wishing to engage in criminal exploitation of children may seek to develop a relationship with one or more minors as part of the grooming process. In such settings, the content may be entirely (or largely) text-based and the collation of evidence against an offender will require the assembly and annotation of such computer-based interactions.

As in the cases of graphic imagery noted earlier, there are several contexts in which exposure to explicit text-based materials may be a professional requirement - as a digital forensic investigator,

a legal professional or a member of law enforcement. With this in mind, our purpose is to explore the possibility of developing techniques, based upon the classification of such textual content, which will provide means to neutralise by varying degrees, the strength of content to which a student, trainee or other professional may be exposed. Our aim is to manage the quantity and degree of such content with a view to minimising any detrimental effects (observer impact) that such content may have on ill-prepared individuals.

2. Observer impact

Anyone who has encountered pornography is likely to be aware of its strong emotional force. Furthermore, this generally increases with the 'strength' of the pornographic content. Although there is some dispute over the long-term psychological impact of such exposure [see, for example, 4, 5, 6], there is strong evidence to suggest that such exposure may detrimentally affect the psychology of the individual. For example, Paolucci et. al. [7] conclude that 'The results are clear and consistent; exposure to pornographic material puts one at increased risk for developing sexually deviant tendencies, committing sexual offenses, experiencing difficulties in one's intimate relationships, and accepting the rape myth' [7]. Other sources echo this perspective for 'strong' pornography: 'Viewers of novel and extreme pornographic images may become tolerant to such images, which may impact sexual response' [8].

Individuals engaged in detection, evidence gathering and legal proceedings involving child pornography, face a risk of 'observer impact', since work commitment means that distancing themselves from such materials becomes impossible. A 2010 study [9] examined the 'psychological impact of viewing disturbing media on investigators engaged in computer forensics work' and concluded that 'substantial percentages of investigators reported poor psychological well-being. Greater exposure to disturbing media was related to higher levels of STSD [secondary

traumatic stress disorder] and cynicism' (p.113). The presence of such 'side-effects' and the need to attend to their impact is clearly noted: 'Although law enforcement agencies need individuals to view potentially disturbing images for investigative purposes, there is little research examining the impact of such work on individuals' psychological well-being. This study was the first to assess quantitatively the levels of burnout and STSD among employees who are required to view disturbing media' (p. 120). In conclusion, this study offered 'empirical data on the poor psychological well-being of law enforcement employees investigating computer child pornography cases in terms of both burnout and secondary traumatic stress' (p. 120).

3. Text-based content

We have seen the development of scales for the 'strength' of pornographic images and these scales have been used in research and for judicial purposes, where their application acknowledges the desire to avoid unnecessary exposure to such content. This focus makes sense, in light of the prevalence of 'strong' imagery and its immediate impact upon an observer. In contrast, sexually explicit text content has not received the same attention. One reason for this disparity may be that textual content is considered less 'forceful' than its graphical counterpart. Nevertheless, there are contexts in which exposure to explicit textual materials may be a professional requirement.

A major concern for law enforcement is the use of social networking and similar chat-based sites for the sexual grooming of minors. While such interactions may facilitate the creation or transfer of pornographic images and lead to child exploitation, the medium of interaction may be entirely (or largely) text. In order to pursue a criminal case against such grooming, law enforcement officers must collate evidence against a suspect. Inevitably, this will require the assembly and annotation of such textual content. Such material is often be highly sexualised and explicit, so we have a context in which several individuals may be exposed to this content, with its associated

'observer impact'. Such a legal case may confront law enforcement officers, digital forensic investigators, legal professionals, and associated support staff, with such materials. This should raise duty of care concerns toward these professionals for whom first-hand exposure to text-based sexually explicit or paedophilic materials is an obligation.

4. Gauging text content

In the research reported here, we have been exploring means to measure and neutralise sexually explicit textual content in the belief that through control of the quantity and degree of such content, there is scope to manage the detrimental effects of observer impact. With this objective in mind, a simple strategy may suggest itself. Since scales for graphical child exploitation materials already exist, could we simply apply COPINE or SAP as the basis for classifying the strength of textual content?

There is a key difference between the intended application of scales like COPINE and SAP on the one hand and scales for grading sexually explicit text on the other. Specifically, the former are directed toward ranking an offence and are brought to bear in cases where mere possession of the imagery may itself be illegal. Hence, application of the imagery scale facilitates judgment on the seriousness of a criminal offence. This has no precise parallel in the context of text content. There is no comparable offence associated with 'possession' of sexually explicit text - to any degree of explicitness.

Still, the descriptions inherent in scales for graphical content could be applied directly to textual content. This would allow the possibility of grading texts according to the content that they describe. A textual description of child exploitation might thereby reflect a high (or low) level on the COPINE scale.

Despite the feasibility of this application, a serious deficiency remains with the notion of applying such a scale as the basis for grading the 'strength' of sexually explicit text. The problem is that COPINE and SAP are solely focused on child exploitation and while this area is most relevant for law enforcement, it excludes many aspects of sexual behaviour and descriptions that have no bearing on underage individuals, but may nonetheless cause concern, anxiety or worse in those obliged to read such content. Evidently, the realm of application for COPINE and SAP does not extend to the full range of content that would reasonably give rise to the duty of care concern, as expressed above.

There are numerous techniques available for automatic text classification and these vary in their approach and effectiveness [10]. Such approaches may be suitable for automatically recognizing that any specific text is sexually explicit. In seeking a means to gauge the degree of sexually explicit content, we have not addressed this identification issue but begin from the assumption that we know the text samples under consideration are sexually explicit to a degree that falls within the realm of concern.

In order to experiment with neutralization of explicit texts, we adopted a simple approach to gauging the degree of explicit text content by accumulating sample texts and creating a dataset of sexually explicit terms. This was gathered through a survey of on-line materials, with initial data extracted from chats logs from the 'perverted justice' web site (www.perverted-justice.com). This was supplemented with a data derived from Internet-sourced sex stories and on-line archives of text-based pornography. On this basis, we derived a list of sexually explicit vocabulary. Using this data set we can match against sample texts in order to quantify the degree of sexually explicit content in the sample.

5. Impact reduction

Our approach to reducing the impact of sexually explicit text is based upon the idea of 'obstructing' the immediate perception of the text content. One benefit of working in the text domain is that such content affords a wider degree of granular analysis and alteration than is possible with images.

For any sample text, we can employ a variety of matching tests to determine what specific lexical components contribute the explicit sexual content. At the simplest level this requires string matching on individual words and multiword units. This can be supplemented with varieties of fuzzy matching, from word stemming, through concept clustering to metaphor matching. In word stemming all variations on a root term are considered equal, e.g. fuck, fucking, fucker, fucked. Stemming simplifies the process of 'weighing' the sexual content, for example, four instances of the same stem rather than four different terms. Concept clustering is another form of aggregation. In this case, instances of the same concepts are treated alike, e.g., fuck, screw, shag, etc. Such clustering also simplifies the counting process. Thereby, we may register three instances of the same concept rather than three different terms. A further approach, proposed but yet to be implemented, employs metaphor matching. In this case, we may selectively match terms that serve as metaphors for sexual features. For example, some references to long pointed instruments can be treated as penis metaphors. Embracing metaphor aims to capture innuendo as well as sexual connotations, such as 'put my sausage in your oven'.

Using such techniques, in combination with an annotated dictionary, we can isolate individual words, multi-word units and larger discourse components and selectively replace these aspects with alternative expressions. This is the basis for our process of neutralization.

Neutralisation is how we term the procedure that gauges and applies a reduction in the emotive force of the sexual content The aim of neutralisation is to modify the content in a manner that can convey its ‘seriousness’ and meaning without full exposure to the ‘raw’ materials. In this regard, we are exploring the feasibility of such neutralisation and strategies for effecting such change in sexually explicit texts. Based upon the estimation of sexually explicit content for any sample text, we aim to apply varying degrees of reduction in explicitness, usually maintaining the meaning but reducing the ‘impact’.

To accommodate better the richness of possible sexual content, we include a further dimension in the estimation. We call this the ‘weight’ of any sexually explicit lexical unit. This denotes the strength of word or expression. Thus, one sexually explicit term may be regarded as stronger than another (even if they have the same meaning). Distinguishing greater or lesser ‘weight’ in source content, allows for selective targeting of expressions with lesser or greater strength. Substitution based on levels of emotive force can be more sophisticated and flexible than a simple binary representation of explicit or not. Furthermore, this flexibility allows for greater variety in handling contexts and individuals for whom neutralisation may be appropriate.

6. Implementation

Following the principles and approach described above, we have implemented a neutralisation test facility that operates with three strength levels and four varieties of item replacement. The item replacement options are:

1. Vowel/asterisk replacement
2. Euphemism
3. Full asterisk replacement
4. Floralisation

Seeking to reduce the immediate impact of the explicit language, we have adopted one technique that is often used for obscuring words, the replacement of characters with asterisks. Two forms of this replacement are used (1 and 3 in our list). In case 1, the first vowel in the sexually explicit expressions is replaced with an asterisk. Displaying explicit texts with such replacements has a minimal effect but still slows down the reader's apprehension of the meaning and thereby reduces its impact. The second replacement strategy replaces sexually explicit terms or expressions with euphemistic alternatives. This is considered more effective in reducing the impact of the original text since there is a greater degree of substitution than with vowel/asterisk replacement. Once again, such replacement aims not obscure the meaning of the original 'raw' text.

A further increase in the level of substitution is achieved by applying full replacement of characters in sexually explicit expressions with asterisks. This has a greater obscuring effect but still tends not to prevent the reader from grasping the intended meaning of the original text (given that the reader appreciates the motivation behind the substitutions). The fourth replacement strategy adopts a radical approach to obscuring expressions in the original explicit texts. Floralisation replaces target terms with the names of fragrant flowers to achieve a different form of cognitive distraction.

In our prototype system, this range of replacement strategies (neutralization options) is combined with our three strength levels to afford a 'replacement matrix'. We can select how any target item is replaced, according to its strength. This permits us to 'mix and match' replacement strategies to accommodate different individuals or different varieties of raw materials.

The system for exploring neutralisation outputs XML marked-up text that indicates the strength of the source item and the type of neutralisation for each change. Figure 2 presents an example

with alternative neutralisation options and their combination. The XML tags indicate the strength level of the targeted item and the replacement technique applied.

Original text

I want to rub your tits and your pussy until my cock explodes.

Asterisk 1

I want to rub your <L3.O2>b*obs</L3.O2> and your <L1.O2>p*ssy until my <L2.O2>c*ck</L2.O2> explodes.

Euphemisms

I want to rub your <L3.O1>breasts</L3.O1> and your <L1.O1>vagina</L1.O1> until my <L2.O1>penis</L2.O1> explodes.

Asterisk 2

I want to rub your <L3.O3>*****</L3.O3> and your <L1.O3>*****</L1.O3> until my <L2.O3>****</L2.O3> explodes.

Floralisation

I want to rub your <L3.O4>blueberry</L3.O4> and your <L1.O4>poppy</L1.O4> until my <L2.O4>campion</L2.O4> explodes.

Mixed (euphemism, floralisation, asterisk 1)

I want to rub your <L3.O2>b*obs</L3.O2> and your <L1.O1>vagina</L1.O1> until my <L2.O4>campion</L2.O4> explodes.

Fig. 2: Example text with alternative replacement strategies

7. Conclusions

In different contexts, there will be a need for different means of quantification. If we are operating with sets of documents, we may be satisfied with a density ratio for each document or an average for documents in the set. If we are tracking dynamic content, such as chat or social media, we may need a dynamic measure of sexually explicit content rather than, or in addition to an aggregated density ratio. In other cases, we may be especially interested in trends.

Documents may be gauged for the density of their sexually explicit content, for example, by an aggregate measure, such as the ratio of sexual content, or by the proportion of sexual content to non-sexual content. A real-time context, such as on-going interactive chat, may require dynamic indicators, such as a sequence of measures for content that facilitate generating a plot for progressive content. Whatever the basis for quantifying the sexually explicit content in text, this may be used as a driver for neutralization.

We have focused attention on the need for quantifying sexual content that is in textual rather than image form and have produced a test bed facility that serves as a means to explore the prospects, techniques and benefits of neutralisation.

References

[1] Quayle, E. The COPINE Project. Irish Probation Journal (Probation Board for Northern Ireland) 5, September, 2008

[2] Taylor, M., Quayle, E., and Holland, G. Child Pornography, the Internet and Offending. The Canadian Journal of Policy Research (ISUMA) 2 (2): 94-100, 2001.

[3] Regina v Oliver. Court of Appeal, 2002. http://www.inquisition21.com/pca_1978/reference/oliver2002.html (Retrieved 10th October, 2014)

[4] Allen, M., D'Alessio, D. A. V. E., & Brezgel, K. (1995). A Meta-Analysis Summarizing the Effects of Pornography II Aggression after Exposure. Human communication research,22 (2), 258-283.

[5] Davies, K. A. (1997). Voluntary exposure to pornography and men's attitudes toward feminism and rape. Journal of Sex Research, 34(2), 131-137.

[6] Malamuth, N. M., Addison, T., & Koss, M. (2000). Pornography and sexual aggression: Are there reliable effects and can we understand them? Annual review of sex research,11 (1), 26-91.

[7] Paolucci, E. O., Genuis, M., & Violato, C. (2000). A meta-analysis of the published research on the effects of pornography. The changing family and child development, 48-59.

[8] Segal, D. (28 March 2014). Does Porn Hurt Children? New York Times. http://www.nytimes.com/2014/03/29/sunday-review/does-porn-hurt-children.html (Retrieved 10th October 2014).

[9] Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. Journal of Police and Criminal Psychology, 25(2), 113-124.

[10] Weir, G. R., & Duta, A. (2012). Strategies for neutralising sexually explicit language. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third* (pp. 66-74). IEEE.